In words, the limiting Hessian matrix is the negative of the limiting information matrix. An analogous result is true for individual observations:

$$E_0\big(D^2_{\theta\theta}\,\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)\big) = -E_0\big(D_\theta^{\top}\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)D_\theta\,\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)\big). \tag{8.44}$$

The latter result clearly implies the former, given the assumptions that permit the application of a law of large numbers to the sequences $\{D^2_{\theta\theta}\,\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)\}_{t=1}^{\infty}$ and $\{D_\theta^{\top}\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)D_\theta\,\ell_t(\boldsymbol{y},\boldsymbol{\theta}_0)\}_{t=1}^{\infty}$.

The result (8.44) is proved by an argument very similar to that used at the beginning of the last section in order to show that the expectation of the CG matrix is zero. From the fact that

$$\frac{\partial \ell_t}{\partial\theta_i} = \frac{1}{L_t}\frac{\partial L_t}{\partial\theta_i},$$

we obtain after a further differentiation that

$$\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j} = \frac{1}{L_t}\frac{\partial^2 L_t}{\partial\theta_i\partial\theta_j} - \frac{1}{L_t^2}\frac{\partial L_t}{\partial\theta_i}\frac{\partial L_t}{\partial\theta_j}.$$

Consequently,

$$\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j} + \frac{\partial \ell_t}{\partial\theta_i}\frac{\partial \ell_t}{\partial\theta_j} = \frac{1}{L_t}\frac{\partial^2 L_t}{\partial\theta_i\partial\theta_j}. \tag{8.45}$$

If now we take the expectation of (8.45) for the DGP characterized by the same value of the parameter vector $\boldsymbol{\theta}$ as that at which the functions ℓ_t and L_t are evaluated (which as usual we denote by E_θ), we find that

$$\begin{aligned} E_\theta\bigg(\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j} + \frac{\partial \ell_t}{\partial\theta_i}\frac{\partial \ell_t}{\partial\theta_j}\bigg) &= \int L_t\frac{1}{L_t}\frac{\partial^2 L_t}{\partial\theta_i\partial\theta_j}\,dy_t \\ &= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int L_t\,dy_t = 0, \end{aligned} \tag{8.46}$$

provided that, as for (8.34), the interchange of the order of differentiation and integration can be justified. The result (8.46) now establishes (8.44), since it implies that

$$E_\theta\bigg(\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j}\bigg) = 0 - E_\theta\bigg(\frac{\partial \ell_t}{\partial\theta_i}\frac{\partial \ell_t}{\partial\theta_j}\bigg) = -E_\theta\bigg(\frac{\partial \ell_t}{\partial\theta_i}\frac{\partial \ell_t}{\partial\theta_j}\bigg).$$

In order to establish (8.43), recall that, from (8.19) and the law of large numbers,

$$\begin{aligned} \mathcal{H}_{ij}(\boldsymbol{\theta}) &= \lim_{n\to\infty}\bigg(\frac{1}{n}\sum_{t=1}^{n}E_\theta\bigg(\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\bigg)\bigg) \\ &= -\lim_{n\to\infty}\bigg(\frac{1}{n}\sum_{t=1}^{n}E_\theta\bigg(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial\theta_j}\bigg)\bigg) \\ &= -\mathcal{I}_{ij}(\boldsymbol{\theta}), \end{aligned}$$

where the last line follows immediately from the definition of the limiting information matrix, (8.22). This then establishes (8.43).

By substituting either $-\mathcal{H}(\boldsymbol{\theta}_0)$ for $\mathfrak{I}(\boldsymbol{\theta}_0)$ or $\mathfrak{I}(\boldsymbol{\theta}_0)$ for $-\mathcal{H}(\boldsymbol{\theta}_0)$ in (8.42), it is now easy to conclude that the asymptotic covariance matrix of the ML estimator is given by either of the two equivalent expressions $-\mathcal{H}(\boldsymbol{\theta}_0)^{-1}$ and $\mathfrak{I}(\boldsymbol{\theta}_0)^{-1}$. Formally, we may write

$$\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \mathfrak{I}^{-1}(\boldsymbol{\theta}_0) = -\mathcal{H}^{-1}(\boldsymbol{\theta}_0).$$

In order to perform any statistical inference, it is necessary to be able to *estimate* $\mathfrak{I}^{-1}(\boldsymbol{\theta}_0)$ or $-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)$. One estimator which suggests itself at once is $\mathfrak{I}^{-1}(\hat{\boldsymbol{\theta}})$, that is, the inverse of the limiting information matrix evaluated at the MLE, $\hat{\boldsymbol{\theta}}$. Notice that the matrix function $\mathfrak{I}(\boldsymbol{\theta})$ is *not* a sample-dependent object. It can, in principle, be computed theoretically as a matrix function of the model parameters from the (sequence of) loglikelihood functions ℓ^n. For some models, this is an entirely feasible computation, and then it yields what is often the preferred estimator of the asymptotic covariance matrix. But for many models the computation, even if feasible, would be excessively laborious, and in these cases it is convenient to have available other consistent estimators of $\mathfrak{I}(\boldsymbol{\theta}_0)$ and consequently of the asymptotic covariance matrix.

One common estimator is the negative of the so-called **empirical Hessian**. This matrix is defined as

$$\hat{\mathcal{H}} \equiv \frac{1}{n}\sum_{t=1}^{n} D^2_{\theta\theta}\,\ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}). \tag{8.47}$$

The consistency of $\hat{\boldsymbol{\theta}}$ and the application of a law of large numbers to the right-hand side guarantees the consistency of (8.47) for $\mathcal{H}(\boldsymbol{\theta}_0)$. When the empirical Hessian is readily available, as it will be if maximization routines that use second derivatives are employed, minus its inverse can provide a very convenient way to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$. However, the Hessian is often difficult to compute, and if it is not already being calculated for other purposes, it probably does not make sense to compute it just to estimate a covariance matrix.

Another commonly used estimator of the information matrix is known as the **outer-product-of-the-gradient estimator**, or **OPG estimator**. It is based on the definition

$$\mathfrak{I}(\boldsymbol{\theta}) \equiv \lim_{n\to\infty}\bigg(\frac{1}{n}\sum_{t=1}^{n} E_{\theta}\big(D_{\theta}^{\top}\ell_t(\boldsymbol{\theta})D_{\theta}\ell_t(\boldsymbol{\theta})\big)\bigg).$$

The OPG estimator is

$$\hat{\mathfrak{I}}_{\mathrm{OPG}} \equiv \frac{1}{n}\sum_{t=1}^{n} D_{\theta}^{\top}\ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}})D_{\theta}\ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{G}^{\top}(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}}), \tag{8.48}$$